TD-MPC2: Scalable, Robust World Models for Continuous Control

Nicklas Hansen[1], Hao Su[1*], Xiaolong Wang[1*]

[1]University of California San Diego

UC San Diego

tdmpc2.com

TD-MPC2

TD-MPC is a model-based reinforcement learning (MBRL) algorithm that performs local trajectory optimization in the latent space of a learned implicit (decoder-free) world model. In this work, we present TD-MPC2: a series of improvements upon the TD-MPC algorithm.
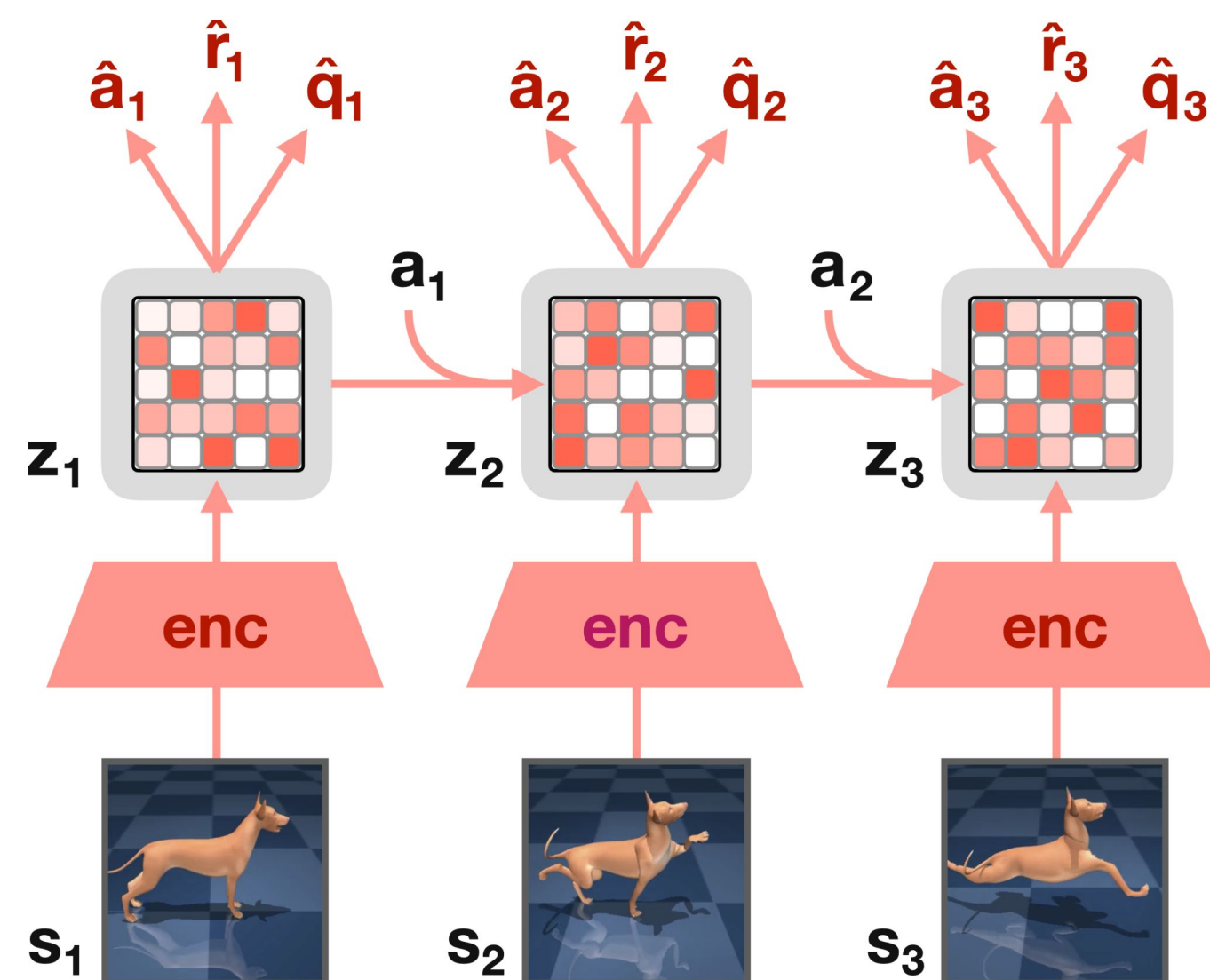
- TD-MPC2 is **robust**: favorable results on 104 online RL tasks *without* hyperparameter-tuning.
- TD-MPC2 is **scalable**: we train a *single* 317M parameter agent to perform 80 tasks.

TD-MPC2 consists of 5 components:

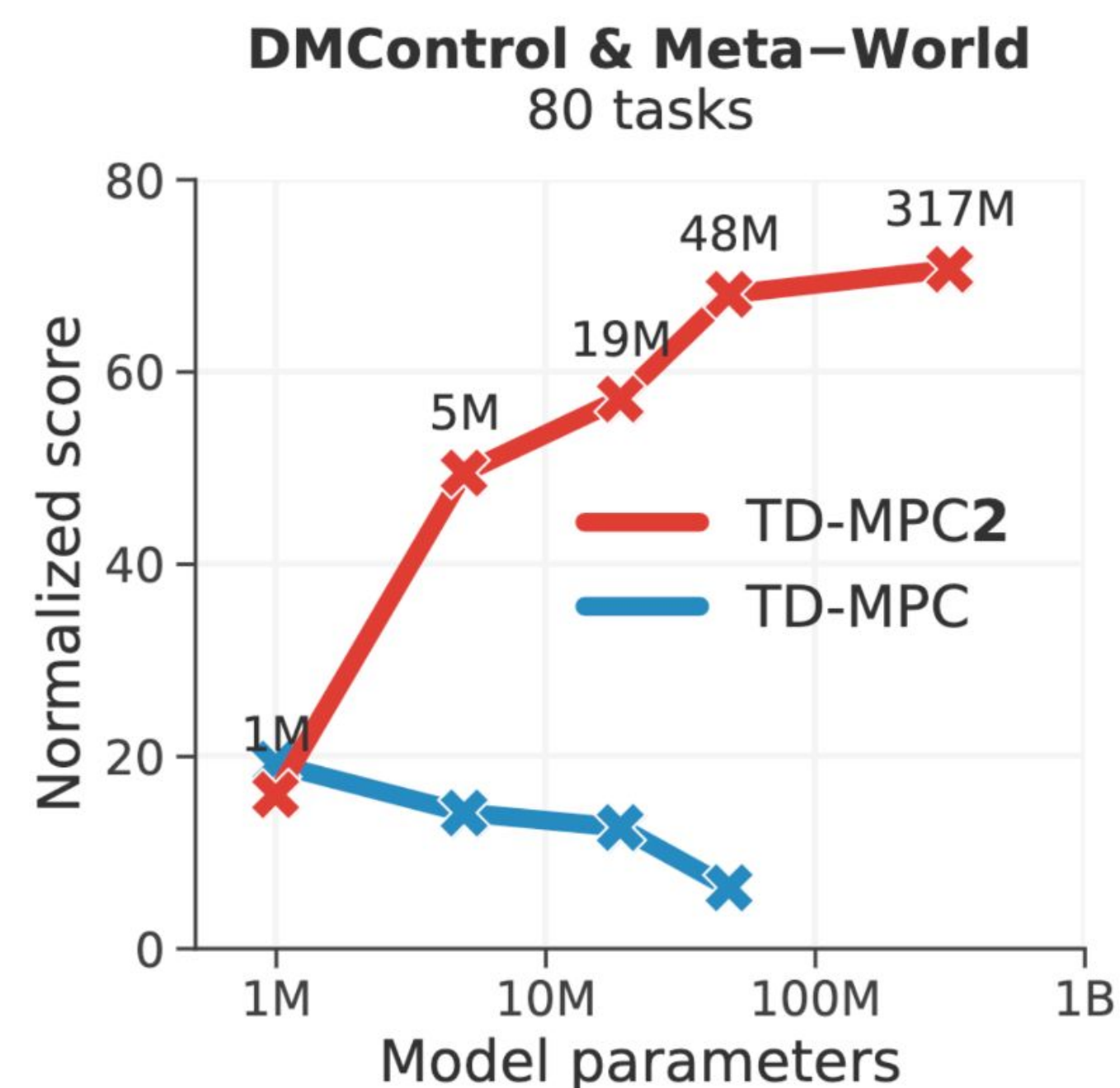| Encoder	$\mathbf{z} = h(\mathbf{s}, \mathbf{e})$
Latent dynamics	$\mathbf{z}' = d(\mathbf{z}, \mathbf{a}, \mathbf{e})$
Reward	$\hat{r} = R(\mathbf{z}, \mathbf{a}, \mathbf{e})$
Terminal value	$\hat{q} = Q(\mathbf{z}, \mathbf{a}, \mathbf{e})$
Policy prior	$\hat{\mathbf{a}} = p(\mathbf{z}, \mathbf{e})$

and is jointly optimized by predicting 3 quantities:

$$\underbrace{\| \mathbf{z}'_t - \mathrm{sg}(h(\mathbf{s}'_t)) \|_2^2}_{\text{Joint-embedding prediction}} + \underbrace{\mathrm{CE}(\hat{r}_t, r_t)}_{\text{Reward prediction}} + \underbrace{\mathrm{CE}(\hat{q}_t, q_t)}_{\text{Value prediction}}$$
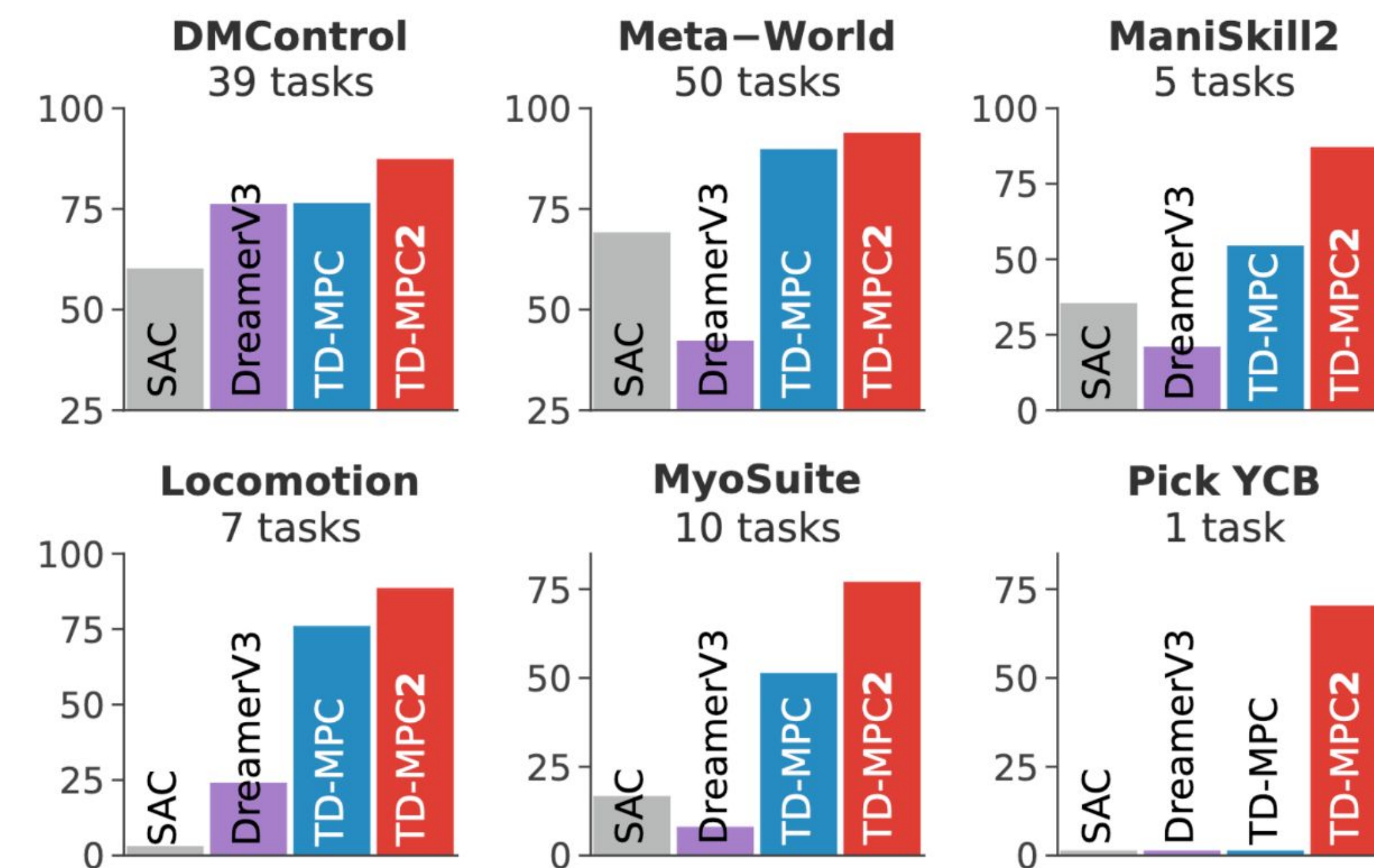

104 continuous control tasks across 4 tasks domains

Supporting open-source science

We release code, **300+** model checkpoints (including 12 multi-task models), and two datasets:

Domains	Tasks	Embodiments	Episodes	Transitions	Size	Link
DMControl + Meta-World	80	12	2.69M	545M	34GB	Download
DMControl	30	11	690k	345M	20GB	Download

Main result – <u>no</u> hyperparameter-tuning!

Multi-task

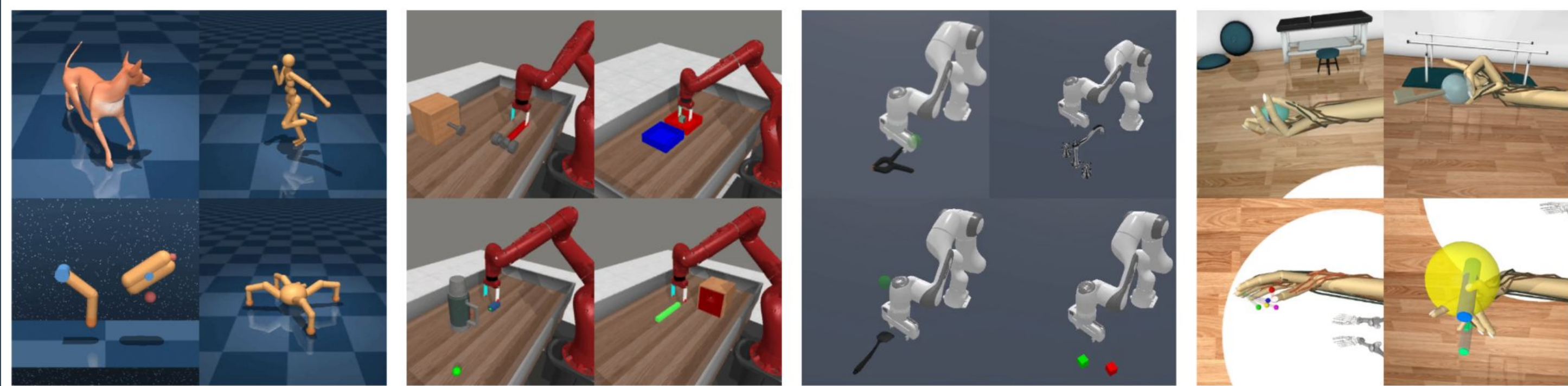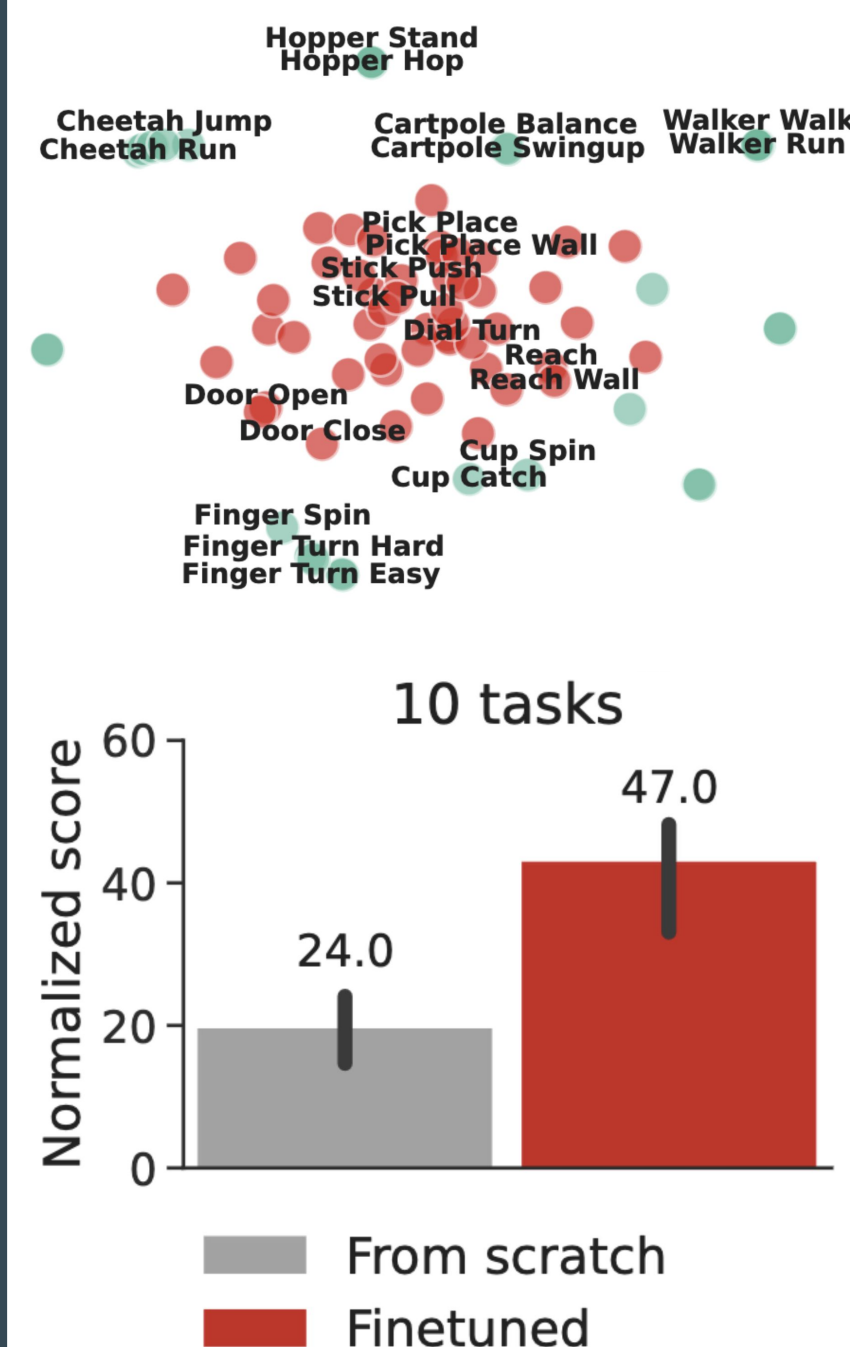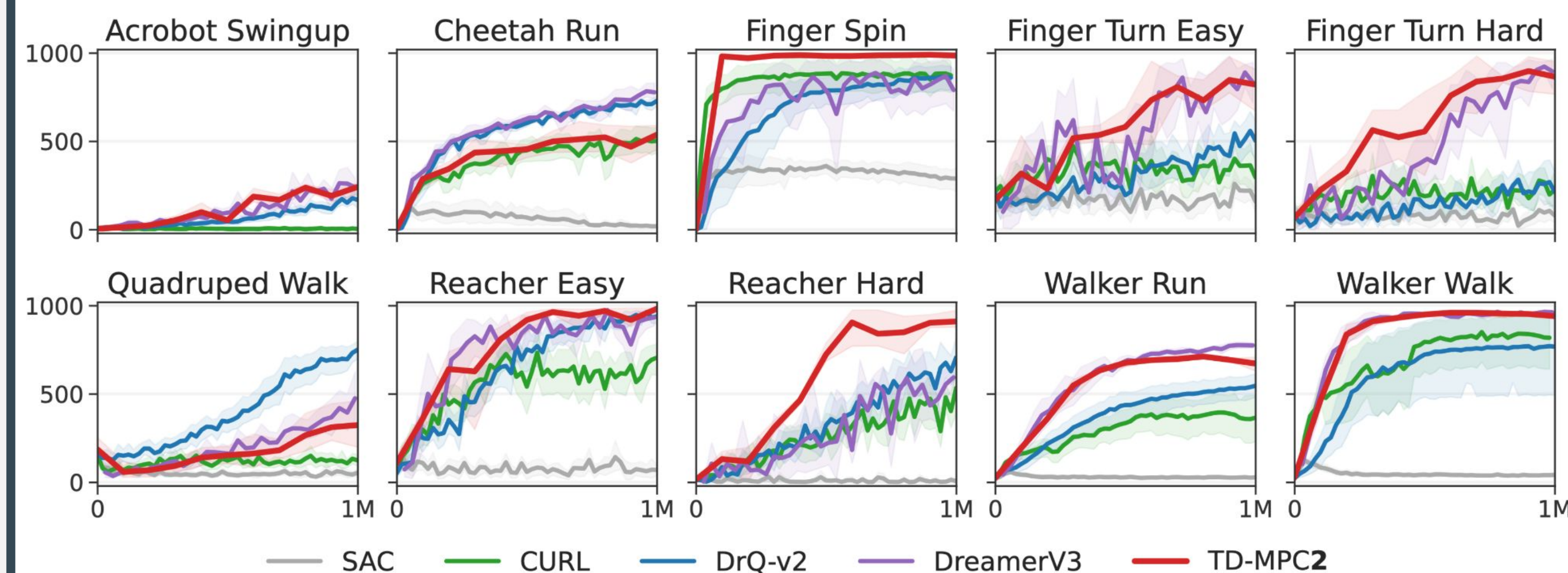**DMControl & Meta-World** — 80 tasks

Single-task

Overview. TD-MPC2 compares favorably to existing model-free and model-based methods across **104** continuous control tasks spanning multiple domains, with a *single* set of hyperparameters (*right*). We further demonstrate the scalability of TD-MPC2 by training a single 317M parameter agent to perform 80 tasks across multiple domains, embodiments, and action spaces (*left*).

Finetuning TD-MPC2

Visual RL

Compares to SOTA algorithms for visual RL – without any tuning! We apply TD-MPC2 to 10 visual RL tasks from DMControl without changing any hyperparameters, and find that performance is comparable to baselines that **do** use task-specific hyperparameters.